Text-based Language Identification of Multilingual Names

Oluwapelumi Giwa and Marelie H. Davel

Abstract:
Text-based language identification (T-LID) of isolated words has been shown to be useful for various speech processing tasks, including pronunciation modelling and data categorisation. When the words to be categorised are proper names, the task becomes more difficult: not only do proper names often have idiosyncratic spellings, they are also often considered to be multilingual. We, therefore, investigate how an existing T-LID technique can be adapted to perform multilingual word classification. That is, given a proper name, which may be either mono- or multilingual, we aim to determine how accurately we can predict how many possible source languages the word has, and what they are. Using a Joint Sequence Modelbased approach to T-LID and the SADE corpus – a newly developed proper names corpus of South African names – we experiment with different approaches to multilingual T-LID. We compare posterior-based and likelihood-based methods and obtain promising results on a challenging task.